AdaBoost

主要内容：

AdaBoost 简介

训练误差分析

一、AdaBoost 简介：

给定训练集：$(x_1, y_1),...,(x_N, y_N)$，其中 $y_i \in \{1,-1\}$，表示 x_i 的正确的类别标签，$i = 1,...,N$

训练集上样本的初始分布：$D_1(i) = \frac{1}{N}$

对 $t = 1,...,T$，

计算弱分类器 $h_t : X \to \{-1,1\}$，该弱分类器在分布 D_t 上的误差为：

$$\varepsilon_t = \mathrm{P}_{D_t}\left(h_t(x_i) \neq y_i\right)$$

计算该弱分类器的权重：$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

更新训练样本的分布：$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$，其中 Z_t 为归一化常数。

最后的强分类器为：

$$H_{final}(x) = sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$$

二、训练误差分析

记 $\varepsilon_t = \frac{1}{2} - \gamma_t$，由于弱分类器的错误率总是比随机猜测（随机猜测的分类器的错误率为 0.5），所以 $\gamma_t > 0$，则训练误差为：

$R_{tr}\left(H_{final}\right) \le \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$。

记 $\forall t, \gamma \ge \gamma_t > 0$，则 $R_{tr}\left(H_{final}\right) \le e^{-2\gamma^2 T}$。

证明：

1、对 D_{T+1} 进行迭代展开

$$D_{T+1}(i) = D_T(i)\frac{\exp\left(-\alpha_T y_i h_T(x_i)\right)}{Z_T}$$

$$= D_1(i)\frac{\exp\left(-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)\right)}{\prod_{t=1}^{T}Z_t} \qquad \text{令 } f(x) = \sum_{t=1}^{T}\alpha_t h_t(x)$$

$$= D_1(i)\frac{\exp\left(-y_i f(x_i)\right)}{\prod_{t=1}^{T}Z_t}。$$

由于 D_{T+1} 是一个分布，

所以：$\sum_{i=1}^{n} D_{T+1}(i) = 1$

所以 $\prod_{t=1}^{T} Z_t = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i f(x_i)\right)$。

2、 训练误差为

$$R_{tr}\left(H_{final}\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\left\{i : y_i \ne H_{final}(x_i)\right\}\right|$$

$$= \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & if\ y_i \ne H_{final}(x_i)\\ 0 & else\end{cases}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 & if\ y_i f(x_i) \le 0\\ 0 & else\end{cases} \qquad *$$

$$\le \frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i f(x_i)\right)$$

$$= \prod_{t=1}^{T} Z_t \text{。}$$

所以 $R_{tr}\left(H_{final}\right) \le \prod_{t=1}^{T} Z_t$，$\prod_{t=1}^{T} Z_t = \frac{1}{N}\sum_{i=1}^{N} \exp\left(-y_i f\left(x_i\right)\right)$ 为训练误差的上界。

相当于损失函数取 $L\left(y, f\left(x\right)\right) = \exp\left(-yf\left(x\right)\right)$，则经验风险/测试误差为 $\frac{1}{N}\sum_{i=1}^{N} \exp\left(-y_i f\left(x_i\right)\right)$，使该经验风险最小的估计为 $f\left(x\right) = \sum_{t=1}^{T} \alpha_t h_t\left(x\right)$。该风险称为指数风险。

*当样本分对时，$y_i f\left(x_i\right) > 0$，所以 $0 < \exp\left(-y_i f\left(x_i\right)\right) < 1$，是一个较小的正数。

当样本分错时，$y_i f\left(x_i\right) \le 0$，所以 $\exp\left(-y_i f\left(x_i\right)\right) \ge 1$。

所以将 $\frac{1}{N}\sum_{i=1}^{N} \begin{cases} 1 & if\ y_i f\left(x_i\right) \le 0 \\ 0 & else \end{cases}$ 变为 $\exp\left(-y_i f\left(x_i\right)\right)$，相当于对上述两种错误率都放大了，这样 $\le$ 不等式成立。

3、 证明 $\alpha_t = \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$；

问题：给定弱分类器的集合：$\Delta = \left\{h_1\left(x\right), h_2\left(x\right), ..., h_M\left(x\right)\right\}$，确定弱分类器 h_t 及其权重 α_t。

$$\left(h, \alpha\right)^* = \arg\min_{\{(\alpha, h)\}} \frac{1}{N}\sum_{i=1}^{N} \exp\left(-y_i f\left(x_i\right)\right)$$

$$= \arg\min \prod_{t=1}^{T} Z_t$$

具体实现时，首先选一个错误率最小的弱分类器 h_t，然后确

定其权重，所以是一个贪心算法。（相当于对 $f(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$，前向逐步递增特征选择，后面再详细描述）

$\frac{\partial Z_t}{\partial \alpha_t}=\frac{\partial\sum_{i=1}^{N}D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{\partial\alpha_t}$，因为 $Z_t=\sum_{i=1}^{N}D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)$

$$=-\sum_{i=1}^{N}D_t(i)\,y_i h_t(x_i)\exp\left(-\alpha_t y_i h_t(x_i)\right)$$

$$=\begin{cases}-\sum_{x_i\in A}D_t(i)\exp(-\alpha_t),if\ \ x_i\in A, A=\{x_i: y_i h_t(x_i)=1\}\\ \sum_{x_i\in\bar{A}}D_t(i)\exp(\alpha_t),if\ \ x_i\in\bar{A}, \bar{A}=\{x_i: y_i h_t(x_i)=-1\}\end{cases}$$

即 A 为分类正确的样本的集合，$\bar{A}$ 为分类错误的样本的集合。

$$\frac{\partial Z_t}{\partial\alpha_t}=0\Rightarrow\sum_{x_i\in A}D_t(i)\exp(-\alpha_t)=\sum_{x_i\in\bar{A}}D_t(i)\exp(\alpha_t)$$

$\Rightarrow\sum_{x_i\in A}D_t(i)\exp(-\alpha_t)=\sum_{x_i\in\bar{A}}D_t(i)\exp(\alpha_t)$，两边同乘以 $\exp(\alpha_t)$

$$\sum_{x_i\in A}D_t(i)=\exp(2\alpha_t)\sum_{x_i\in\bar{A}}D_t(i)$$

正确率=$\sum_{x_i\in A}D_t(i)=1-\varepsilon_t$，错误率=$\sum_{x_i\in\bar{A}}D_t(i)=\varepsilon_t$，

所以 $1-\varepsilon_t=\varepsilon_t\exp(2\alpha_t)$

所以 $\alpha_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$。

当 ε_t 很小时，α_t 很大，即错误率很小的弱分类器的权重很大。

4、 训练误差

$$Z_t=\sum_{i=1}^{N}D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)$$

$$=\sum_{x_i\in A}D_t(i)\exp(-\alpha_t)+\sum_{x_i\in\bar{A}}D_t(i)\exp(\alpha_t)$$

$$= -(1-\varepsilon_t)\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} + \varepsilon_t\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$$

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

令 $\varepsilon_t = \frac{1}{2} - \gamma_t$，由于弱分类器的错误率总是比随机猜测（随机猜测的分类器的错误率为 0.5），所以 $0 < \gamma_t < \frac{1}{2}$，

所以 $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = 2\sqrt{\left(\frac{1}{2}-\gamma_t\right)\left(1-\frac{1}{2}+\gamma_t\right)} = 2\sqrt{\left(\frac{1}{2}-\gamma_t\right)\left(\frac{1}{2}+\gamma_t\right)}$

$= \sqrt{1-4\gamma_t^2} \le \sqrt{e^{-4\gamma_t^2}} = e^{-2\gamma_t^2}$ （不等式可利用 e^{-x^2} 在 $x=0$ 处 Taylor 展开得到）

令 $\forall t, \gamma \ge \gamma_t > 0$，即 γ 为所有 γ_t 中最小的一个。

则训练误差的上界为：

$$R_{tr}\left(H_{final}\right) \le \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} e^{-2\gamma_t^2} \le \prod_{t=1}^{T} e^{-2\gamma^2} = \exp\left(-2\sum_{t=1}^{T}\gamma^2\right) = e^{-2\gamma^2 T}。$$

所以，当 $T \to \infty, R_{tr}\left(H_{final}\right) \le e^{-2\gamma^2 T} \to 0$，即训练误差的上界随 T 的增加指数减小。

5、 AdaBoost 相当于最大贝叶斯后验

$$(h,\alpha)^* = \arg\min_{\{(\alpha,h)\}_1^M} \frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i f(x_i)\right)，$$

当损失函数取 $L\left(y, f(x)\right) = \exp\left(-yf(x)\right)$ 时，则上述表达式为经验风险，当样本很多时，样本均值趋近于期望，即期望风险/测试误差为，

$$E = \sum_{i=1}^{N} p(x_i, y_i)\exp\left(-y_i f(x_i)\right)，$$ （$p(x_i, y_i)$ 表示 (x_i, y_i) 的概率密度函数）

$$= \sum_{y_i=1} p(y_i=1)\, p(x_i \mid y_i=1) \exp\left(-f(x_i)\right) + \sum_{y_i=-1} p(y_i=-1)\, p(x_i \mid y_i=-1) \exp\left(f(x_i)\right)$$

$$= p(y=1)\, p(x \mid y=1) \exp\left(-f(x)\right) + p(y=-1)\, p(x \mid y=-1) \exp\left(f(x)\right)$$

我们目标是风险最小的 $f(x)$，即

$$\frac{\partial E}{\partial f} = -p(y=1)\, p(x \mid y=1) \exp\left(-f(x)\right) + p(y=-1)\, p(x \mid y=-1) \exp\left(f(x)\right)$$

$$= -p(y=1)\, p(x \mid y=1) + p(y=-1)\, p(x \mid y=-1) \exp\left(2f(x)\right)$$
$$= 0$$

所以

$$f(x) = \frac{1}{2} \log \frac{p(y=1)\, p(x \mid y=1)}{p(y=-1)\, p(x \mid y=-1)} = \frac{1}{2} \log \frac{p(y=1 \mid x)}{p(y=-1 \mid x)}$$

所以

$$H(x) = sign\left(f(x)\right) = sign\left(\log \frac{p(y=1 \mid x)}{p(y=-1 \mid x)}\right),$$

为最大贝叶斯后验。

上面证明了收敛性，最后的强分类器收敛于最大后验概率。

6、 AdaBoost 相当于前向逐步递增加法建模

$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$，

可视为基展开，其中 $h_t(x)$ 为基函数，α_t 为对应基函数的权重。对基展开，通常是给定基函数，一次联合求出所有的基函数中的参

数及其权重α_t（如用最小二乘法或极大似然估计方法）。

而 AdaBoost 为一个逐步递增的方式增加基函数，并计算其权重，不调整已添加的基函数中的参数及其权重。

假设第$T-1$步的模型为： $f_{T-1}(x)=\sum_{t=1}^{T-1}\alpha_t h_t(x)$

当损失函数取$L(y,f(x))=\exp(-yf(x))$时，则第 T 步新增加的基函数h_T及其权重α_T要使得训练误差/经验风险最小，即

$$(h_T,\alpha_T)=\arg\min_{h,\alpha}\sum_{i=1}^{N}\exp\left(-y_i\left(f_{T-1}(x_i)+\alpha h(x_i)\right)\right),$$

$$=\arg\min_{h,\alpha}\sum_{i=1}^{n}w_i^T\exp\left(-\alpha y_i h(x_i)\right),$$

其中$w_i^T=\exp(-y_i f_{T-1}(x_i))$。因为每个$w_i^T$不依赖于$\alpha,h$，所以$w_i^T$可以看作是应用于每个观测的权值，该权值依赖于$f_{T-1}(x_i)$，所以，每个样本的权值随每次迭代改变。

上述问题可以分两步实现：

第一步：首先选一个错误率最小的弱分类器h_T，

$$h_T=\arg\min_{h}\sum_{i=1}^{N}w_i^T I\left(y_i\neq h(x_i)\right)。$$

第二步：然后确定其权重α_T

因为

$$\sum_{i=1}^{N}w_i^T\exp\left(-\alpha y_i h(x_i)\right)=e^{-\alpha}\sum_{y_i=h(x_i)}w_i^T+e^{\alpha}\sum_{y_i\neq h(x_i)}w_i^T$$

$$=\left(e^{\alpha}-e^{-\alpha}\right)\sum_{i=1}^{N}w_i^T I\left(y_i\neq h(x)\right)+e^{-\alpha}\sum_{i=1}^{N}w_i^T$$

将h_T代入，即可得到

$$\alpha_T = \frac{1}{2}\ln\left(\frac{1-\varepsilon_T}{\varepsilon_T}\right),$$

其中ε_T表示错误率。